Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte,
Carlos Ramisch, Uxoa Iñurrieta, Voula Giouli

Accepted for publication in the Prague Bulletin of Mathematica Linguistics
https://ufal.mff.cuni.cz/pbml/112

June 13, 2019

Multiword expressions are. . .

Definition

Combinations of words which exhibits lexical, morphosyntactic, semantic, pragmatic and/or statistical idiosyncrasies

Characteristics

- Discontinuous → *Carlos* ***made*** *an unusual* ***presentation***
- Non compositional → a ***hot dog*** is not a *dog*
- Ambiguous → a ***piece of cake*** is something easy or something to eat
- . . .

Multiword expressions are. . .

Definition

Combinations of words which exhibits lexical, morphosyntactic, semantic, pragmatic and/or statistical idiosyncrasies

Characteristics

- Discontinuous → *Carlos* ***made*** *an unusual* ***presentation***
- Non compositional → a ***hot dog*** is not a *dog*
- Ambiguous → a ***piece of cake*** is something easy or something to eat
- . . .

This presentation is about MWE ambiguity

MWEs can have literal occurrences

(1) The boss was **pulling** the **strings** from prison. (EN)
'The boss was making use of his influence while in prison.'

(2) You control the marionette by pulling the strings. (EN)

But what is a literal occurrence?

(3) As an effect of pulling, the strings broke. (EN)

(4) He strings paper lanterns on trees without pulling the table. (EN)

(5) Determine the maximum force you can pull on the string so that the string does not break. (EN)

(6) My husband says no **strings** were **pulled** for him. (EN)

(7) She moved Bill by **pulling** wires and **strings**. (EN)

(8) The article addresses the **strings** which the journalist claimed that the senator **pulled**. (EN)

(9) The strings pulled the bridge. (EN)

(10) He was there, **pulling** the **strings**, literally and metaphorically. (EN)

Three research questions

1. How to **define** literal occurrences of MWEs?
2. How **frequent** are literal occurrences of MWEs?
   - Should MWE identification systems take ambiguity into account?
   - Should downstream NLP applications care about them?
3. What are the **cross-lingual** characteristics of literal occurrences?
   - Study them in Basque, German, Greek, Polish and Portuguese

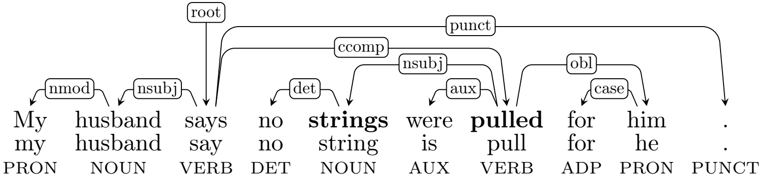
Context

Focus on **verbal multiword expressions** (VMWEs) in the **PARSEME corpora** using **Universal Dependencies** as syntactic formalism

Outline

Outline

Sequence



Sequence

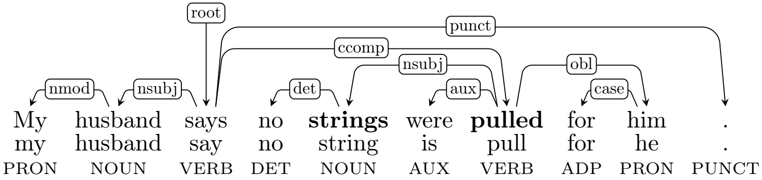
A sentence is viewed as a sequence $s : \{1, 2, \ldots, |s|\} \rightarrow W$

W is the set of all possible word forms (including punctuation)

Equivalently: $s = \{s_1, s_2, \ldots, s_{|s|}\} = \{(1, w_1), (2, w_2), \ldots, (|s|, w_{|s|})\}$

Example: $s = \{(1, \text{My}), (2, \text{husband}), (3, \text{says}), \ldots, (9, \text{him}), (10, .)\}$

Subsequence



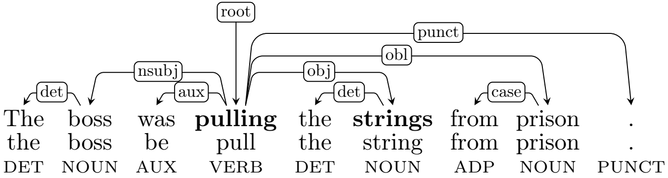
Subsequence

p subsequence of s iff there is an injection $\text{sub}_p^s : \{1, \ldots, |p|\} \rightarrow \{1, \ldots, |s|\}$:

1. $\forall i \in \{1, 2, \ldots |p|\},\ p(i) = s(\text{sub}_p^s(i))$
2. $\forall i, j \in \{1, 2, \ldots |p|\}$, if $i < j$, then $\text{sub}_p^s(i) < \text{sub}_p^s(j)$.

Example: $p = \{p_1, p_2\} = \{(1, \text{strings}), (2, \text{pulled})\}$ $\text{sub}_p^s(1) = 5$ and $\text{sub}_p^s(2) = 7$

Dependency graph

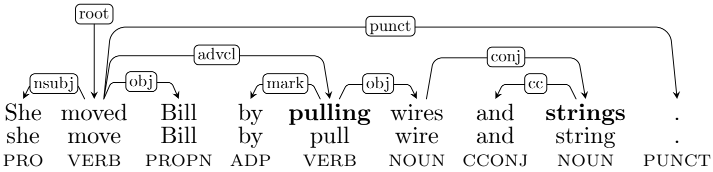


Dependency graph

A dependency graph of a sequence s is a tuple $\langle V_s, E_s \rangle$:

- $V_s = \{\langle 1, \text{surface}(s_1), \text{lemma}(s_1), \text{pos}(s_1)\rangle, \ldots, \langle |s|, \text{surface}(s_{|s|}), \text{lemma}(s_{|s|}), \text{pos}(s_{|s|})\rangle\}$
- E_s is the set of labeled edges connecting nodes in V_s

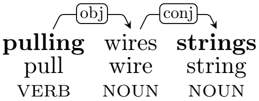
Example: $\text{label}(s_2) = \text{nsubj}$, $\text{parent}(s_2) = s_4$, $\text{label}(s_4) = \text{root}$, $\text{parent}(s_4) = \text{nil}$



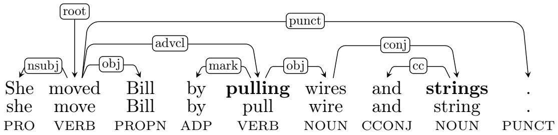
Dependency subgraph

A dependency subgraph $\langle V_p, E_p \rangle$ is a minimal weakly connected graph[a] containing *at least* the nodes corresponding to p.

[a] Connected, ignoring the directions of edges.

Example: 

Coarse syntactic structure (CSS)

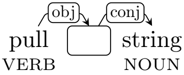


Coarse syntactic structure (CSS)

The coarse syntactic structure $\mathrm{css}(p) = \langle V_{\mathrm{css}(p)}, E_{\mathrm{css}(p)} \rangle$ of a subsequence p is a directed graph:

- $V_{\mathrm{css}(p)} = \{\langle \mathrm{lemma}(p_1), \mathrm{pos}(p_1)\rangle, \ldots, \langle \mathrm{lemma}(p_{|p|}), \mathrm{pos}(p_p)\rangle\}_{\mathrm{ms}} \cup \{D_1, \ldots, D_k\}$
  D_i are dummy nodes replacing the intervening words
- $E_{\mathrm{css}(p)} = E_p$

Example:



VMWE token

VMWE token

A VMWE token e is a subsequence of a sentence s:

1. e has at least **two words**, that is, $|e| > 1$
2. all components $e_1, \ldots, e_{|e|}$ are **lexicalized**[a]
3. the head of each of e's canonical forms must be a **verb**
4. $css(e)$ has no dummy nodes, i.e. e yields a **weakly connected graph**
5. e in s must have an **idiomatic meaning** (e.g. using PARSEME tests)

[a] If they are absent, the VMWE looses the idiomatic meaning.

Canonical form, canonical structure

Canonical form

A canonical form is a minimal VMWE token in its least marked form:

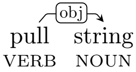
- Finite verb, active voice (if possible)
- No extraction, relative clause, negation (if possible)
- Singular nouns (if possible)

Example: *he* ***pulled*** *the* ***strings***

Canonical structure

The canonical structure of a VMWE is the coarse syntactic structure (CSS) of its canonical forms

Example:



VMWE type

VMWE variant set

A VMWE variant set is an (infinite) set of VMWE tokens sharing the **same CSS** and the same meaning.

Example: {*he **pulled** the **strings**, we **pull** some **strings**, ...*}

VMWE type

A VMWE type is an (infinite) set of VMWE variant sets sharing the same set of **CSS vertices** and the same meaning.

Example: {*he **pulled** the **strings**, we **pull** some **strings**, ...*}
∪ { *no **strings** were **pulled**, many **strings** are **pulled**...*}
∪ ...

Idiomatic, literal and coincidental occurrences I

- s is a sentence of length $|s|$
- t is a VMWE type $t = \{\langle css_1, \sigma_{ID} \rangle, \ldots, \langle css_{|t|}, \sigma_{ID} \rangle\}$, $css_i = \langle V, E_i \rangle$
- A potential occurrence p of t in s is a subsequence of s, $V_{css(p)} = V$

Idiomatic, literal and coincidental occurrences II

Idiomatic occurrence (IO)

1. The CSS of p is identical to one of the CSSes in t
2. p occurs with the meaning σ_{ID}

Literal occurrence (LO)

1. There is a rephrasing s' of s (possibly identical) such that:
   1. s' is synonymous with s
   2. there is a subsequence p' in s' such that $V_{\text{css}(p)} = V_{\text{css}(p')}$
   3. the CSS of p' is equal to the canonical structure of t
2. p does not occur with the meaning σ_{ID}

Coincidental occurrence (CO)

- there is no rephrasing s' of s which fulfills conditions (1-3) of an LO.

Applying the definitions

(11)	The boss was **pulling** the **strings** from prison.	(IO)
(12)	You control the marionette by pulling the strings.	(LO)
(13)	As an effect of pulling, the strings broke.	(CO)
(14)	He strings paper lanterns on trees without pulling the table.	(None)
(15)	The force you can pull on the string so that it does not break.	(CO)
(16)	My husband says no **strings** were **pulled** for him.	(IO)
(17)	She moved Bill by **pulling** wires and **strings**.	(IO)
(18)	The **strings** which he claimed that the senator **pulled**.	(IO)
(19)	The strings pulled the bridge.	(CO)
(20)	He was there, **pulling** the **strings**, literally and metaphorically.	(?)

Outline

Corpus

- PARSEME shared task v1.1 corpora
- Manual annotation for VMWE tokens:
  - Inherently reflexive verbs (IRV)
  - Ligt-verb constructions (LVC)
  - Verb-particle constructions (VPC)
  - Verbal idioms (VID)
- Manual or automatic lemmas, UD POS tags, UD morphological features, UD dependency trees

Corpus stats

Lang.	Sent.	Tokens	VMWEs	Morphology	Syntax
Basque	11,158	157,807	3,823	partly manual	partly manual
German	8,996	173,293	3,823	automatic	automatic
Greek	8,250	224,762	2,405	automatic	automatic
Polish	16,121	274,318	5,152	partly manual	partly manual
Portuguese	27,904	638,002	5,536	partly manual	partly manual

Outline

Relaxed non idiomatic occurrences (RNOs)

Goal: extract potential LOs from the corpus for annotation

Procedure

1. extract each VMWE token $e = \{e_1, \ldots, e_{|e|}\}$ in each sentence s
2. for each extracted e, for each sentence $s' = \{s'_1, s'_2, \ldots, s'_{|s'|}\}$:
3. r is a relaxed non-idiomatic occurrence (RNO) of e in s', if:
   - r is a subsequence of s'
   - $|r| = |e|$
   - there is a bijection $\text{rno}_e^r : \{1, \ldots, |e|\} \rightarrow \{1, \ldots, |e|\}$ such that:
     - for $i \in \{1, 2, \ldots, |e|\}$ and $j = \text{rno}_e^r(i)$, $cf(\text{lemmasurface}(e_i)) \in \{cf(\text{lemma}(r_j)), cf(\text{surface}(r_j))\}$
     - r is not a VMWE token

LO candidates

- **WindowGap**: all matched tokens of the RNO must fit into a sliding window with no more than g external elements (gaps). We use $g = 2$.
- **BagOfDeps**: the RNO must corresponding to a weakly connected unlabeled subgraph with no dummy nodes
- **Unlabeled**: the RNO must correspond to a connected unlabeled graph with no dummy nodes, that is, the dependency labels are ignored but the parent relations are preserved.
- **Labeled**: the RNO must be a connected labeled graph with no dummy nodes, in which both the parent relations and the dependency labels are preserved.

The resulting set of LO candidates is the **union of the 4 heuristics** output

Outline

First phase: initial checks

- $e = \{e_1, e_2, \ldots, e_{|e|}\}$ is a VMWE token annotated in a sentence s
- cs is the canonical structure of e's type
- $c = \{c_1, c_2, \ldots, c_{|c|}\}$ is an LO candidate extracted by the heuristics

1. [**FALSE**] Should e have been annotated as an IO of an MWE at all?
   - NO → annotate c as ERR-FALSE-IDIOMATIC
   - YES → go to the next test
2. [**SKIP**] Is c an IO of an MWE that annotators forgot/ignored?
   - YES, it is a verbal MWE → annotate c as ERR-SKIPPED-IDIOMATIC
   - YES, but a non-verbal MWE → annotate c as NONVERBAL-IDIOMATIC
   - UNSURE, not enough context → annotate c as MISSING-CONTEXT
   - NO → go to the next test
3. [**LEX**] Do c's components have the same lemma and POS as cs's?
   - NO → annotate c as WRONG-LEXEMES
   - YES → go to the next test

Second phase: classification

1. **[COINCIDENCE]** Are the syntactic dependencies in c equivalent to those in cs? Dependencies are considered equivalent if a rephrasing (possibly identical) of s is possible, keeping its original sense and producing dependencies identical to those in cs.
   - NO → annotate c as COINCIDENTAL
   - YES → go to the next test
2. **[MORPH]** Could the knowledge of morphological constraints allow us to automatically classify c as an LO?
   - YES → annotate c as LITERAL-MORPH
   - NO or UNSURE → go to the next test
3. **[SYNT]** Could the knowledge of syntactic constraints allow us to automatically classify c as an LO?
   - YES → annotate c as LITERAL-SYNT
   - NO or UNSURE → annotate c as LITERAL-OTHER

Examples

- ERR-FALSE-IDIOMATIC:
  - *She [. . .] brought back a branch of dill.*
- ERR-SKIPPED-IDIOMATIC:
  - *Bring down* in *Any insult [. . .]* ***brings*** *us all* ***down***
- NONVERBAL-IDIOMATIC:
  - *After the major* ***kill-offs****, wolves [. . .]*
- MISSING-CONTEXT:
  - *Enron is blowing up.*
- WRONG-LEXEMES:
  - *Then take your finger and place it under their belly*
- COINCIDENTAL: (***do the job***)
  - *[. . .] why you like the job and do a little bit of [. . .]*
- LITERAL-MORPH: (***get going***)
  - *At least you get to go to Florida [. . .]*
- LITERAL-SYNT: (***have to do*** *with something*)
  - *[. . .] we have better things to do.*
- LITERAL-OTHER: (***come of it***)
  - *[. . .] we've come out of it quite good friends*

Known limitations

Syntactic framework (UD) can change annotation

- *the* ***presentation*** *was* ***made***
- *his presentation made a good impression*
- *we made a surprise at her presentation*

Granularity of relations can change the annotation

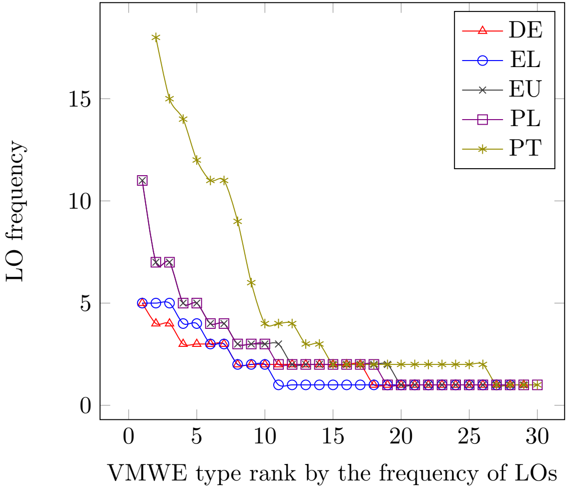
- Reflexive clitics annotated as `expl` with "semantic" subrelations

Outline

Overall results

	DE		EL		EU		PL		PT	
Annotated IOs	3,823		2,405		3,823		4,843		5,536	
LO candidates	926		451		2,618		332		1,997	
ERR-FALSE-ID.	21.5%	(199)	12.0%	(54)	9.4%	(246)	0.0%	(0)	3.8%	(76)
ERR-SKIPPED-ID.	27.0%	(250)	47.5%	(214)	17.3%	(453)	5.4%	(18)	10.7%	(213)
NONVERBAL-ID.	0.0%	(0)	0.0%	(0)	0.2%	(6)	0.0%	(0)	0.5%	(9)
MISSING-CONTEXT	0.3%	(3)	0.2%	(1)	0.5%	(12)	2.1%	(7)	0.7%	(13)
WRONG-LEXEMES	40.1%	(371)	0.9%	(4)	26.7%	(700)	1.8%	(6)	38.1%	(760)
COINCIDENTAL (COs)	**2.6%**	(24)	**27.9%**	(126)	**42.4%**	(1110)	**61.1%**	(203)	**33.5%**	(668)
LITERAL (LOs)	**8.5%**	(79)	**11.5%**	(52)	**3.5%**	(91)	**29.5%**	(98)	**12.9%**	(258)
↪ LITERAL-MORPH	0.8%	(7)	5.5%	(25)	1.9%	(51)	1.2%	(4)	3.7%	(73)
↪ LITERAL-SYNT	1.5%	(14)	2.0%	(9)	0.7%	(19)	8.1%	(27)	2.2%	(44)
↪ LITERAL-OTHER	6.3%	(58)	4.0%	(18)	0.8%	(21)	20.2%	(67)	7.1%	(141)
Idiomaticity rate	**98%**		**98%**		**98%**		**98%**		**96%**	

Distribution of LOs



Performance of the heuristics

Language	WindowGap			BagOfDeps			Unlabeled			Labeled			All (union)		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Basque	3	**91**	7	6	89	**11**	5	58	9	**6**	22	10	3	100	7
German	8	78	14	12	**90**	22	13	**90**	22	**14**	77	**23**	9	100	16
Greek	11	87	20	15	**90**	26	**16**	83	**27**	16	52	24	12	100	21
Polish	33	**96**	49	43	81	56	49	73	**59**	**52**	23	32	30	100	46
Portuguese	14	**98**	25	17	62	27	20	59	30	**34**	37	**36**	13	100	23

Outline

LOs of IRVs

(21)	Nesse	rio	se	encontraram	muitos	tipos	de	peixe. (PT)
	In.this	river	RCLI	found/met	many	kinds	of	fish.

‘Many kinds of fish were found in this river.’

Finding: Some IRVs are ambiguous with middle-passive and impersonal

LOs of LVCs

(22)	Nie	**mają**	wymaganego	**zezwolenia**	na	pracę. (PL)
	Not	have.3rd.PL	required	permission	for	work.

‘They have no permission to work.’

(23)	Kierowcy	mieli	sfałszowane	zezwolenia.	(PL)
	Drivers	had	falsified	permissions.	

‘The drivers had false driving licenses.’

Finding: LOs of LVCs occur when the predicative noun is polysemous

Portuguese-specific LVC LOs I

Resultatives:

(24)	Ele	tem	sua	força	renovada	quando	descansa.	(PT)
	He	has	his	strength	renewed	when	rests.	

'His strength gets renewed when he rests.'

(25)	A	criança	**tem**	uma	**alimentação**	equilibrada.	(PT)
	The	child	has	a	diet	balanced.	

'The child has a balanced diet.'

Portuguese-specific LVC LOs II

Secondary predication:

(26) João tem [seu irmão]$_{obj}$ [como um demônio]$_{iobj}$. (PT)
John has his brother as a demon.
'João considers his brother a demon.'

(27) Eles **tem** [essa atividade]$_{obj}$ [**como** uma **opção**]$_{iobj}$. (PT)
they have this activity as an option.
'This activity is a possible option for them.'

Finding: some language-specific phenomena require syntactic constraints to distinguish LOs from IOs

LOs of VIDs I

(28) (EU)

Gaixo	dago	eta	ez	**da**	joateko	**gauza**.
Sick	is	and	no	is	going	thing

He/She is sick and is no thing to go.
'He/She is sick and is unable to go.'

(29) (EU)

Horiek	beste	garai	bat-eko	gauza-k	dira.
These	other	time	one-GEN	thing-PL	AUX

These are things from the past.
'These things belong to the past.'

Finding: many VID LOs can be identified with morphological constraints

LOs of VIDs II

(30) Służenie nam **mają** **we krwi**. (PL)
serving us have.3rd.PL in blood
They have serving us in blood.
'Serving us is their innate ability.'

(31) Miał we krwi ponad 1,5 promila alkoholu (PL)
had.3rd.SING in blood over 1.5 per-mille alcohol
'His blood alcohol level was 1.5.'

Finding: domain-specific uses can be LOs of general-purpose IOs

Basque-specific VID LOs I

(32) **Kontu-a-n** **hartu** du lagun-a-ren
account-ART-LOC take AUX friend-ART-GEN
iritzi-a. (EU)
opinion-ART.ABS

Took into account the opinion of his/her friend.
'He/She took his/her friend's opinion into account.'

(33) Diru-a hartu du kontu-tik. (EU)
money-ART.ABS take AUX account-ABL

Took money from the account.
'He/She withdraw money from the account.'

Finding: lemmas + POS in CSSes inadequate for agglutinative languages

Outline

Take-home message

1. Good parsers (taggers, etc.) are required to distinguish IOs from COs
2. LOs are theoretically possible, but not so frequent in practice (2-4%)
3. Simple heuristics + special cases could identify most VMWE IOs
4. Do we need machine learning to identify known VMWEs?
5. What kind of constraints need to be encoded in lexicons? And how?
6. Can these constraints be discovered using semi-supervised learning?

Take-home message

1. Good parsers (taggers, etc.) are required to distinguish IOs from COs
2. LOs are theoretically possible, but not so frequent in practice (2-4%)
3. Simple heuristics + special cases could identify most VMWE IOs
4. Do we need machine learning to identify known VMWEs?
5. What kind of constraints need to be encoded in lexicons? And how?
6. Can these constraints be discovered using semi-supervised learning?

Literal occurrences of VMWEs are **rare birds** that **cause a stir**